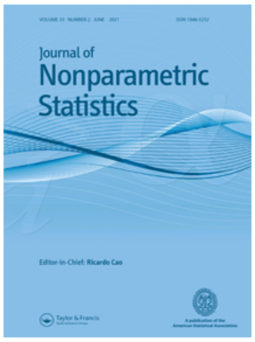

Taylor & Francis
Taylor & Francis Group

Journal of Nonparametric Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gnst20

Truncation data analysis for the under-reporting probability in COVID-19 pandemic

Wei Liang, Hongsheng Dai & Marialuisa Restaino

To cite this article: Wei Liang, Hongsheng Dai & Marialuisa Restaino (2021): Truncation data analysis for the under-reporting probability in COVID-19 pandemic, Journal of Nonparametric Statistics, DOI: 10.1080/10485252.2021.1989426

To link to this article: https://doi.org/10.1080/10485252.2021.1989426

JOURNAL OF NONPARAMETRIC STATISTICS
https://doi.org/10.1080/10485252.2021.1989426

Taylor & Francis
Taylor & Francis Group

Truncation data analysis for the under-reporting probability in COVID-19 pandemic

Wei Liang[a], Hongsheng Dai[b] and Marialuisa Restaino[c]

[a]Xiamen University, Xiamen, People's Republic of China; [b]Mathematical Sciences, University of Essex, Colchester, UK; [c]Department of Economics and Statistics, University of Salerno, Salerno, Italy

ABSTRACT
The COVID-19 pandemic has affected all countries in the world and brings a major disruption in our daily lives. Estimation of the prevalence and contagiousness of COVID-19 infections may be challenging due to under-reporting of infected cases. For a better understanding of such pandemic in its early stages, it is crucial to take into consideration unreported infections. In this study we propose a truncation model to estimate the under-reporting probabilities for infected cases. Hypothesis testing on the differences in truncation probabilities, that are related to the under-reporting rates, is implemented. Large sample results of the hypothesis test are presented theoretically and by means of simulation studies. We also apply the methodology to COVID-19 data in certain countries, where under-reporting probabilities are expected to be high.

ARTICLE HISTORY
Received 1 October 2020
Accepted 30 September 2021

KEYWORDS
COVID-19; hypothesis test; large sample theory; truncation; under-reporting

AMS SUBJECT CLASSIFICATIONS
62G10; 62G20

1. Introduction

The COVID-19 pandemic has progressively expanded to almost every country in the world. By September 2020, the new virus SARS-CoV-2 has infected more than 33 million people and the death toll is more than 1 million. To model the epidemic for such a highly infectious disease, Susceptible-Infectious-Recovery (SIR) models are widely used. Estimating the model parameters, the infection rate, recovery rate, death rate, etc. in the early stage is very important to help government better control the epidemic and minimise its impacts on our society. Although there is a vast literature related to COVID-19 modelling and estimation of SIR model parameters (Huang et al. 2020; Zhang 2020; Zhou et al. 2020; Zhu et al. 2020), their results may not be reliable because of the under-reporting, which leads to selection bias in the dataset collected in the pandemic.

Under-reporting of infected cases and deaths has resulted in delayed responses in many countries, which cause severe social and economic impacts. Under-reporting is due to, for instance, people who actually died from the virus before they were diagnosed to be infected or people having mild symptoms and recovered from it before they even realised that they had been infected already. Therefore existing research outputs may very likely estimate

CONTACT Hongsheng Dai hdaia@essex.ac.uk

Supplemental data for this article can be accessed here. https://doi.org/10.1080/10485252.2021.1989426

the death rate or infection rate with bias. For example, in the UK, many old people in care homes died before they were diagnosed and these death cases were not included in the early stage of the pandemic. Also there were adjustments on the total number of infected cases in China, UK and France, at later stages when more and more under-reporting evidence had been gathered.

Under-reporting in the early stages of the pandemic makes it very difficult to understand the epidemic (Maugeri, Barchitta, Battiato, and Agodi 2020; Nishiura et al. 2020). Although existing research has pointed out that unrecognised cases (often patients who experienced mild or no symptoms, and patients that died before hospitalisation) could silently expose a far greater proportion of the population to the SARS-CoV-2 (Li et al. 2020), they mainly used simple estimation approaches or simulation studies to provide a subjective under-reporting probability estimate.

In this research, we propose to use a truncation model to estimate the under-reporting probabilities for infected cases. If the track-and-trace can be implemented perfectly in each country, then every patient will have a reporting time X^* (time from having symptoms to reporting date). However, this is not the case in practice. The reporting time X^* can only be recorded if the patient has been hospitalised (or tested due to any reason). In other words, we can only observe X^* if $X^* \geq Y^*$, where Y^* is the time from the date having symptoms to the date that patients go to hospital (or being tested). If $X^* \geq Y^*$, we observe both X^* and Y^*; otherwise we do not have any information of X^* and Y^*. Note that X^* may be also subject to right censoring by variable C^*, which is usually the last follow-up date, for example from symptom date to recovery or from symptom date to other events which result reporting not happening yet.

Such truncation and censoring models have been proven useful in many areas including cancer research, clinical trials, epidemiological studies and actuarial science (Lawless 2003; Kalbfleisch and Prentice 2011). We here focus on hypothesis testing to study the under-reporting rate for infected cases under different characteristic (such as age and country) groups.

Our research proposes a new hypothesis testing method to study the difference of truncation probabilities in different population groups, where the truncation probability links to the under-reporting probability. We prove the large sample results for the hypothesis test statistic, under the martingale framework. Note that, existing research in the literature mainly focussed on comparison between survival curves of different populations, for example the Gehan test and log-rank test (Lagakos, Barraj, and De Gruttola 1988; Bilker and Wang 1996; Shen 2007, 2009, 2015). We use the nonparametric product limit estimator to estimate the unknown survival curve in our analysis. Based on the estimated survival functions, the hypothesis test statistic (for truncation probabilities) is constructed. Although nonparametric maximum likelihood estimator (NPMLE) is another option for estimating the survival function, it may underestimate the survival function at early times for small to medium sized samples in the presence of truncation (Lynden-Bell 1971; Woodroofe 1985; Tsai 1988).

This paper is organised as follows. The methodologies are presented in Section 2, including the hypotheses, test statistic and its large sample results, with all necessary proofs provided in Appendix. Simulation studies and data analysis are presented in Sections 3 and 4, respectively. The paper ends with a conclusion in Section 5.

2. Methodology

2.1. Model framework and the test hypotheses

Suppose there are K populations. For each $k = 1, 2, \dots, K$, denote $(X^*_{i,k}, Y^*_{i,k}, C^*_{i,k})$, $i = 1, 2, \dots$ as the continuous random variables from the kth population such that $X^*_{i,k}$ is independent of $(Y^*_{i,k}, C^*_{i,k})$. The main survival time of interest $X^*_{i,k}$ (time from the date having symptoms to the reporting date) is subject to left truncation by $Y^*_{i,k}$ (time from the date having symptoms to the date being referred to hospital) and is also subject to right censoring by $C^*_{i,k}$. This means we observe nothing if $X^*_{i,k} < Y^*_{i,k}$, and observe $\tilde{X}^*_{i,k} = \min\{X^*_{i,k}, C^*_{i,k}\}$ and $\delta^*_{i,k} = I[X^*_{i,k} \le C^*_{i,k}]$ if $X^*_{i,k} \ge Y^*_{i,k}$. We also assume $Y^*_{i,k}$ is independent of $C^*_{i,k}$ and $\mathbf{P}(C^*_{i,k} > Y^*_{i,k}) = 1$ throughout this paper, for notation simplicity. Note that the independent assumption of Y^* and C^* can be removed and the corresponding theoretical proofs are provided in the Supplementary file.

Denote the survival function, the cumulative distribution and the cumulative hazard function for $X^*_{i,k}$ as $S_k(t)$, $F_k(t) = 1 - S_k(t) = \mathbf{P}(X^*_{1,k} \le t)$ and $\Lambda_k(t)$, respectively. We further denote the cumulative distribution for $Y^*_{i,k}$ and $C^*_{i,k}$ as $G_k(t) = \mathbf{P}(Y^*_{1,k} \le t)$ and $Q_k(t) = \mathbf{P}(C^*_{1,k} \le t)$, and the cumulative hazard function for $Y^*_{i,k}$ as $\Lambda_{k,G}(t)$.

For any cumulative distribution function $F(x)$, denote

$$a_F = \inf\{x : F(x) > 0\}, \quad \text{and} \quad b_F = \sup\{x : F(x) < 1\}.$$

We impose the following condition throughout the paper.

Condition 2.1: For $k = 1, 2, \dots, K$, $F_k(\cdot)$, $G_k(\cdot)$ and $Q_k(\cdot)$ are continuous in their support $[a_{F_k}, b_{F_k}]$, $[a_{G_k}, b_{G_k}]$ and $[a_{Q_k}, b_{Q_k}]$, respectively.

Further, from Woodroofe (1985), under the following condition, F_k, G_k and Q_k are identifiable. Denote $\tau_k = \min\{b_{F_k}, b_{Q_k}\}$.

Condition 2.2: For $k = 1, 2, \dots, K$,

$$a_{G_k} \le \min\{a_{F_k}, a_{Q_k}\}, \quad \text{and} \quad b_{G_k} \le \tau_k = \min\{b_{F_k}, b_{Q_k}\}.$$

The boundary assumption $b_{G_k} \le \min(b_{F_k}, b_{Q_k})$ includes the case $b_{G_k} = \min(b_{F_k}, b_{Q_k}) = \infty$ which is true in the early stage of the pandemic. This is because at the early stage of the pandemic cases will only be recorded if patients are referred to hospital and then tested. We may have $b_{G_k} = \infty$ (implying some patients being infected have mild symptoms and they never go to hospital), but for such patients their reporting times can also be defined as ∞, i.e. $b_{F_k} = \infty$. Since we can also assume $b_{Q_k} = \infty$ for such patients because of the nonparametric analysis used, the boundary assumption $\min(b_{F_k}, b_{Q_k}) \le b_{G_k}$ is still valid.

The truncation probability for group k, $k = 1, 2, \dots, K$, is $\alpha_k = \mathbf{P}(X^*_{i,k} \ge Y^*_{i,k}) = \int G_k(s)\, \mathrm{d}F_k(s)$. The probability α_k represents the probability that a subject $X^*_{i,k}$ can be observed, and therefore $1 - \alpha_k$ is the under-reporting probability for the kth population. Our main research target is to test the hypotheses

$$H_0 : \alpha_1 = \cdots = \alpha_K \leftrightarrow H_1 : \alpha_1 \ge \cdots \ge \alpha_K \tag{1}$$

with at least one $\ge$ to be strictly $>$.

2.2. The test statistic

To develop a test statistic for (1), we need to introduce the following notations first. Denote the observed biased sample for the kth population as

$$(Y_{1,k}, \tilde{X}_{1,k}, \delta_{1,k}), (Y_{2,k}, \tilde{X}_{2,k}, \delta_{2,k}), \ldots, (Y_{n_k,k}, \tilde{X}_{n_k,k}, \delta_{n_k,k}),$$

where $\tilde{X}_{i,k} = \min\{X_{i,k}, C_{i,k}\}$ and $\delta_{i,k} = I[X_{i,k} \leq C_{i,k}]$. Denote $n = \sum_{k=1}^{K} n_k$. Note that throughout this paper, notations with superscript $*$, such as $X^*_{i,k}$ and $Y^*_{i,k}$ mean the random variables from the population without selection bias, however such values cannot be observed; and notations without superscript $*$, such as $\tilde{X}_{i,k}$ and $Y_{i,k}$, mean the observed truncated samples (having selection bias).[1]

Define a counting process related to $\tilde{X}_{i,k}$

$$\mathrm{d}N_{i,k}(t) = I[t \leq \tilde{X}_{i,k} < t + \mathrm{d}t, \delta_{i,k} = 1], \quad N_k(t) = \sum_{i=1}^{n_k} N_{i,k}(t), \tag{2}$$

and

$$H_{i,k}(t) = I[\tilde{X}_{i,k} \geq t > Y_{i,k}], \quad \bar{H}_k(t) = \sum_{i=1}^{n_k} H_{i,k}(t). \tag{3}$$

Then, the cumulative hazard function of $X^*_{i,k}$ can be estimated via

$$\mathrm{d}\hat{\Lambda}_k(t) = \frac{\mathrm{d}N_k(t)}{\bar{H}_k(t)},$$

and its survival function $S_k(t) = 1 - F_k(t)$ can be estimated by the product limit estimator

$$\hat{S}_k(t) = \begin{cases} 1, & 0 \leq t < \tilde{X}_{(1),k}, \\ \prod\limits_{s \in (0,t]} \left[1 - \dfrac{\mathrm{d}N_k(s)}{\bar{H}_k(s)}\right], & \tilde{X}_{(1),k} \leq t < \tilde{X}_{(n_k),k}, \\ 0, & t \geq \tilde{X}_{(n_k),k}, \end{cases} \tag{4}$$

where $\tilde{X}_{(1),k} = \min\{\tilde{X}_{i,k}\}$, $\tilde{X}_{(n_k),k} = \max\{\tilde{X}_{i,k}\}$. From Woodroofe (1985), we know that the estimate of $G_k(t)$ is defined by

$$\hat{G}_k(t) = \begin{cases} 0, & 0 \leq t < Y_{(1),k}, \\ \prod\limits_{s > t} \left[1 - \dfrac{\mathrm{d}N_{k,G}(s)}{\bar{H}_k(s)}\right], & Y_{(1),k} \leq t < Y_{(n_k),k}, \\ 1, & t \geq Y_{(n_k),k}, \end{cases} \tag{5}$$

where $N_{k,G}(t)$ is given by

$$\mathrm{d}N_{k,G}(t) = \sum_{i=1}^{n_k} I[t \leq Y_{i,k} < t + \mathrm{d}t],$$

and $Y_{(1),k} = \min\{Y_{i,k}\}$, $Y_{(n_k),k} = \max\{Y_{i,k}\}$. The large sample properties of $\hat{\Lambda}_k$ and $\hat{S}_k$ are provided in Appendix 1.

For $k = 1, 2, \ldots, K$, denote

$$\hat{\alpha}_k = \int \hat{G}_k(s)\,\mathrm{d}\hat{F}_k(s) = \int \hat{G}_k(s)\hat{S}_k(s)\,\mathrm{d}\hat{\Lambda}_k(s). \tag{6}$$

Then, the test statistic can be constructed as

$$\begin{aligned} W &= \sum_{k=1}^{K-1}\left(\hat{\alpha}_k - \frac{\hat{\alpha}_{k+1} + \cdots + \hat{\alpha}_K}{K-k}\right) = \sum_{k=1}^{K-1}\hat{\alpha}_k - \sum_{k=1}^{K-1}\sum_{i\geq k+1}\frac{\hat{\alpha}_i}{K-k} \\ &= \sum_{k=1}^{K-1}\hat{\alpha}_k - \sum_{i=2}^{K}\left(\sum_{k=1}^{i-1}\frac{1}{K-k}\right)\hat{\alpha}_i \\ &= \hat{\alpha}_1 + \sum_{k=2}^{K-1}\left(1 - \sum_{i=1}^{k-1}\frac{1}{K-i}\right)\hat{\alpha}_k - \left(\sum_{i=1}^{K-1}\frac{1}{K-i}\right)\hat{\alpha}_K. \end{aligned}$$

Let $c_1 = 1$, $c_k = 1 - \sum_{i=1}^{k-1}(K-i)^{-1}$, $k = 2, \ldots, K-1$ and $c_K = -\sum_{i=1}^{K-1}(K-i)^{-1}$, then W can be simplified as

$$W = \sum_{k=1}^{K} c_k\hat{\alpha}_k = \sum_{k=1}^{K} c_k \int \hat{G}_k(s)\,\mathrm{d}\hat{F}_k(s). \tag{7}$$

Note that $\sum_{k=1}^{K} c_k = 0$.

2.3. The large sample properties of the test statistic

The following two assumptions are needed to ensure that the variance of the test statistic exists.

Assumption 2.1: For $k = 1, 2, \ldots, K$, $\Lambda_k(t) < \infty$, for $t \in [0, \tau_k)$, and

$$\int_0^{\tau_k} \frac{\mathrm{d}\Lambda_k(s)}{G_k(s)(1 - Q_k(s))S_k(s)} < \infty.$$

Assumption 2.2: For $k = 1, 2, \ldots, K$, $\Lambda_{k,G}(t) < \infty$, for $t \in [0, \tau_k)$, and

$$\int_0^{\tau_k} \frac{\mathrm{d}\Lambda_{k,G}(s)}{G_k^2(s)(1 - Q_k(s))^2 S_k^2(s)} < \infty.$$

Theorem 2.1: *Suppose for $k = 1, 2, \ldots, K$, $n_k/n \to p_k \in (0, 1)$, and Conditions 2.1, 2.2 and Assumptions 2.1, 2.2 hold. Denote $L_k(t) = \mathbf{P}(\tilde{X}_{i,k} \geq t > Y_{i,k})$. Under H_0, as $n \to \infty$, we* *have*

$$\sqrt{n}\,W \to N\left(0, \sigma_W^2\right),$$

where

$$\sigma_W^2 = \sum_{k=1}^{K} p_k^{-1} c_k^2 \sigma_k^2, \quad \sigma_k^2 = \sigma_{k1}^2 + \sigma_{k2}^2 + \sigma_{k3}^2,$$

and

$$\sigma_{k1}^2 = \int_0^{\tau_k} \left(\int_s^{\tau_k} S_k(u)\,\mathrm{d}G_k(u) \right)^2 \frac{\mathrm{d}\Lambda_k(u)}{L_k(u)},$$

$$\sigma_{k2}^2 = \int_0^{\tau_k} \left(\int_0^{s} G_k(u)\,\mathrm{d}S_k(u) \right)^2 \frac{\mathrm{d}G_k(u)}{\alpha_k L_k^2(u)},$$

$$\sigma_{k3}^2 = 2\int_0^{\tau_k} \left(\int_0^{s} \frac{\int_0^u G_k(v)\,\mathrm{d}S_k(v)}{L_k(u)}\,\mathrm{d}\Lambda_{k,G}(u) \right) \frac{\int_0^s G_k(v)\,\mathrm{d}S_k(v)}{G_k^2(s)}\,\mathrm{d}G_k(s).$$

Theorem 2.1 shows that under H_0, the test statistic $\sqrt{n}W$ converges to a Normal distribution. The rejection region for testing H_0 at the significance level b (setting as $b = 0.05$ in this paper) is

$$\{\sqrt{n}W/\sigma_W \geq z_b\},$$

where z_b is the b upper quantile of $N(0, 1)$. In application, we can use a Bootstrap method to get an consistent estimator $\hat{\sigma}_W$, replacing of σ_W.

3. Simulation studies

We implement a simulation study in this section to examine the finite sample performance of the test statistic W in terms of type I errors. Here we consider two scenarios. In scenario 1, we simulate X_k^*, C_k^* and Y_k^* from normal distribution, while in scenario 2, we consider Weibull distribution. In all of our simulation, we set $K = 3$ groups.

Scenario 1, we set that the null hypothesis H_0 holds, i.e. $H_0 : \alpha_1 = \alpha_2 = \alpha_3$ and check the probability of falsely rejecting H_0 in our simulations. For $i = 1, 2, \ldots, n$, we generate the pseudo data $X_{i,k}^*$, $C_{i,k}^*$ and $Y_{i,k}^*$ from N$(8.5 + k, 1)$, N$(\mu_C + k, 1)$ and N$(\mu_Y + k, 1)$, respectively. The simulated data $\tilde{X}_{i,k}^* = \min\{X_{i,k}^*, C_{i,k}^*\}$, $Y_{i,k}^*$ and $\delta_{i,k}^* = \min\{X_{i,k}^*, C_{i,k}^*\}$ satisfying the truncation condition $X_{i,k}^* > Y_{i,k}^*$ will be used in our study; observations not satisfying the truncation condition will be discarded. The parameter μ_C takes values 10.8, 10 and 9.5 which will give censoring proportions in different groups as about 5%, 15% and 25%, respectively. The truncation parameter μ_Y takes values 6.3, 7 and 7.5 which will make the truncation probabilities of α_k as about 95%, 85% and 75%, respectively. The sample sizes n are chosen as 100, 200, 400 and 600, and the simulation replication is 5000 times. The significance level is chosen as $b = 0.05$.

The results are shown in Table 1. We can see that $\hat{\alpha}_k$ becomes closer to the true value of α_k (truncation proportion), as n increases. Rejection probability for H_0 also goes to the 5% significant level (the type I error being allowed), as n increases.

Table 1. Simulation results*(100) when the null hypothesis holds under normal distributions.

			$n = 100$		$n = 200$		$n = 400$		$n = 600$	
TP	CP	k	$\hat{\alpha}_k$	std.	$\hat{\alpha}_k$	std.	$\hat{\alpha}_k$	std.	$\hat{\alpha}_k$	std.
		1	94.97	0.05	94.70	0.02	94.49	0.01	94.41	0.01
	5%	2	94.97	0.05	94.69	0.03	94.48	0.01	94.35	0.01
		3	95.00	0.05	94.67	0.03	94.47	0.01	94.36	0.01
		RP	3.06		3.88		4.40		4.13	
		1	95.04	0.05	94.69	0.03	94.49	0.01	94.39	0.01
95%	15%	2	95.01	0.05	94.71	0.02	94.49	0.01	94.43	0.01
		3	95.03	0.05	94.76	0.03	94.51	0.01	94.39	0.01
		RP	2.64		3.42		4.18		4.45	
		1	95.09	0.05	94.83	0.03	94.55	0.01	94.44	0.01
	25%	2	95.07	0.05	94.79	0.03	94.54	0.01	94.43	0.01
		3	95.14	0.05	94.82	0.03	94.58	0.01	94.44	0.01
		RP	2.62		3.22		3.65		4.65	
		1	87.54	0.23	86.82	0.12	86.51	0.06	86.36	0.04
	5%	2	87.47	0.23	86.95	0.11	86.54	0.06	86.34	0.04
		3	87.54	0.21	87.00	0.11	86.47	0.06	86.32	0.04
		RP	3.74		3.64		4.88		5.00	
		1	87.71	0.21	87.11	0.11	86.66	0.06	86.48	0.04
85%	15%	2	87.62	0.22	87.15	0.11	86.63	0.06	86.48	0.04
		3	87.67	0.22	87.14	0.11	86.59	0.06	86.41	0.04
		RP	3.70		4.24		4.25		4.80	
		1	87.89	0.22	87.30	0.12	86.85	0.06	86.57	0.04
	25%	2	87.85	0.24	87.30	0.12	86.82	0.06	86.53	0.04
		3	87.90	0.21	87.25	0.12	86.77	0.06	86.59	0.04
		RP	3.30		4.48		4.40		4.65	
		1	79.17	0.49	78.24	0.27	77.47	0.15	77.32	0.10
	5%	2	79.15	0.49	78.29	0.26	77.50	0.15	77.31	0.10
		3	79.30	0.48	78.27	0.26	77.47	0.15	77.33	0.09
		RP	3.92		4.34		4.05		4.48	
		1	79.70	0.49	78.55	0.28	77.83	0.15	77.51	0.10
75%	15%	2	79.74	0.49	78.53	0.28	77.83	0.15	77.49	0.11
		3	79.45	0.52	78.72	0.27	77.93	0.14	77.52	0.11
		RP	4.16		4.00		4.28		4.58	
		1	79.89	0.54	78.96	0.28	78.96	0.28	77.82	0.10
	25%	2	80.17	0.50	78.97	0.27	78.97	0.27	77.76	0.11
		3	79.91	0.54	79.11	0.29	79.11	0.29	77.81	0.10
		RP	4.06		4.24		4.38		4.80	

Note: $\hat{\alpha}_k$ – the mean estimate of the 5000 simulation replicates.std – the standard deviation of $\hat{\alpha}_k$.TP – truncation proportion; $\alpha := \alpha_1 = \cdots = \alpha_K$.CP – censoring proportion.RP – probability of rejecting H_0, of the 5000 simulations.

Scenario 2, we still set that the null hypothesis H_0 holds, and consider the simulated data $X^*_{i,k}$ such that $X^*_{i,k} - k$ from Weibull$(6.5, 4)$, where 6.5 is the scale parameter and shape parameter is 4. The censored variables $C^*_{i,k}$ is such that $C^*_{i,k} - k \sim$ Weibull$(s_C, 4)$, while the truncation variable values $Y^*_{i,k}$ is such that $Y^*_{i,k} - k \sim$ Weibull$(s_Y, 4)$. Since the simulated probability of rejecting H_0 is not sensitive to censoring proportion, we set s_C to be 10 so that the censoring proportion is about 15%. In the meantime, we set s_Y to be 3, 4.1, 4.8 to test the influence of different truncation probabilities on hypothesis testing. The same as above, the sample sizes are chosen as 100, 200, 400 and 600, and the simulation replication is 5000 times. The significance level is chosen as $b = 0.05$. The results are listed in Table 2. We can see that the results for Scenario 2 (Weibull distribution) are very similar to that of Scenario 1, i.e. the estimator $\hat{\alpha}_k$ converges to the true truncation proportion and the convergence becomes faster when the truncation proportion becomes smaller.

Table 2. Simulation results*(100) when the null hypothesis holds under Weibull distributions.

			$n = 100$		$n = 200$		$n = 400$		$n = 600$	
CP	TP	k	$\hat{\alpha}_k$	std.	$\hat{\alpha}_k$	std.	$\hat{\alpha}_k$	std.	$\hat{\alpha}_k$	std.
		1	95.86	0.14	95.75	0.09	95.73	0.05	95.73	0.03
	95%	2	95.99	0.12	95.79	0.08	95.73	0.05	95.72	0.03
		3	95.93	0.13	95.75	0.10	95.73	0.05	95.73	0.03
		RP	1.82		1.92		2.56		2.90	
		1	87.29	0.40	86.88	0.25	86.62	0.16	86.56	0.11
15%	85%	2	87.50	0.40	86.84	0.26	86.73	0.14	86.61	0.12
		3	87.17	0.46	87.01	0.25	86.61	0.16	86.58	0.09
		RP	3.06		3.10		3.50		3.56	
		1	78.99	0.71	78.43	0.39	77.87	0.24	77.71	0.17
	75%	2	79.30	0.71	78.24	0.40	77.79	0.25	77.69	0.18
		3	79.37	0.69	78.14	0.45	77.89	0.23	77.76	0.16
		RP	3.66		4.12		4.20		4.21	

Table 3. Simulation results*(100) when the null hypothesis is false; PoT – power of test.

				$n = 50$		$n = 100$		$n = 150$		$n = 200$	
	TP	CP	k	$\hat{\alpha}_k$	std.	$\hat{\alpha}_k$	std.	$\hat{\alpha}_k$	std.	$\hat{\alpha}_k$	std.
$X \sim$ N(9+0.25k,1)	$\alpha_1 = 81.14\%$	3%	1	84.77	0.63	83.51	0.33	83.13	0.22	82.84	0.18
$Y \sim$ N(7+k,1)	$\alpha_2 = 64.12\%$	1%	2	70.31	1.70	68.69	0.98	67.46	0.69	66.93	0.52
$C \sim$ N(11+k,1)	$\alpha_3 = 43.04\%$	1%	3	55.97	3.37	51.28	1.93	50.03	1.40	48.53	1.08
			PoT	33.23		61.73		78.88		89.58	
$X \sim$ N(9+0.5k,1)	$\alpha_1 = 85.43\%$	4%	1	88.36	0.40	87.57	0.21	87.05	0.16	86.90	0.12
$Y \sim$ N(7+k,1)	$\alpha_2 = 75.98\%$	2%	2	80.53	0.90	79.09	0.48	78.27	0.35	78.36	0.26
$C \sim$ N(11+k,1)	$\alpha_3 = 63.87\%$	1%	3	70.64	1.71	68.59	0.96	67.55	0.70	67.04	0.51
			PoT	25.58		50.08		66.88		80.15	
$X \sim$ N(9+0.75k,1)	$\alpha_1 = 89.40\%$	6%	1	91.45	0.25	90.83	0.13	90.53	0.09	90.24	0.07
$Y \sim$ N(7+k,1)	$\alpha_2 = 85.51\%$	4%	2	88.27	0.40	87.43	0.20	87.16	0.16	86.93	0.11
$C \sim$ N(11+k,1)	$\alpha_3 = 81.13\%$	3%	3	84.84	0.62	83.49	0.37	83.23	0.21	82.84	0.18
			PoT	10.55		21.98		32.68		42.20	
$X \sim$ N(9+0.25k,1)	$\alpha_1 = 80.55\%$	11%	1	84.76	0.62	83.82	0.32	83.32	0.22	83.13	0.18
$Y \sim$ N(7+k,1)	$\alpha_2 = 63.68\%$	4%	2	70.89	1.69	68.54	0.95	67.97	0.64	67.18	0.54
$C \sim$ N(10+k,1)	$\alpha_3 = 43.15\%$	2%	3	56.17	3.33	51.75	1.96	49.60	1.39	48.75	1.11
			PoT	32.70		61.75		79.95		89.03	
$X \sim$ N(9+0.5k,1)	$\alpha_1 = 84.98\%$	15%	1	88.50	0.41	87.76	0.21	87.31	0.15	87.09	0.11
$Y \sim$ N(7+k,1)	$\alpha_2 = 75.47\%$	8%	2	80.63	0.94	79.38	0.48	78.61	0.34	78.34	0.27
$C \sim$ N(10+k,1)	$\alpha_3 = 63.76\%$	4%	3	70.77	1.69	68.71	0.98	67.93	0.68	67.41	0.53
			PoT	25.10		50.40		66.60		79.15	
$X \sim$ N(9+0.75k,1)	$\alpha_1 = 88.35\%$	19%	1	91.76	0.24	90.90	0.13	90.59	0.08	90.39	0.07
$Y \sim$ N(7+k,1)	$\alpha_2 = 84.84\%$	15%	2	88.60	0.39	87.59	0.21	87.30	0.14	87.14	0.11
$C \sim$ N(10+k,1)	$\alpha_3 = 80.59\%$	11%	3	85.03	0.62	83.80	0.32	83.32	0.22	82.96	0.17
			PoT	10.10		21.93		31.83		41.38	

To justify the performance of our method, we implement another simulation to study the power of the test. **Scenario 3**, random variables are simulated from different normal distribution settings which are detailed in Table 3. The truncation probabilities are different in different groups, i.e. H_1 holds. The sample sizes are chosen as 50, 100, 150, and 200 and 5000 simulations are implemented. The significance level is also chosen as $b = 0.05$.

From the results in Table 3, we can see that the estimates for $\hat{\alpha}_k$ converges to the true values α_k as n increases. Also the probability of correctly rejecting H_0 goes higher as n increases. This indeed implies that the power of the test increases as n increases. The greater the difference of truncation proportions between different groups, the higher the power of the test. The censoring proportion has little effect on the power of this test. Note that, in

Table 3 the results of tests of rows 1, 2, 4, 5 have larger powers because these scenarios have very different $\alpha_k, k = 1, 2, 3$ and the hypothesis test method has larger power to distinguish them even with small sample sizes. On the contrary, the results in rows 3 and 6 have smaller power, because in these scenarios the truncation probabilities in different groups $\alpha_k,\ k = 1, 2, 3$ have very close values. It needs much larger sample sizes to distinguish them, which is indeed what we shall expect for all hypothesis test methods.

4. Real data analysis

We study a dataset collected in the early stage of the COVID-19 outbreak, from January to February, 2020, provided in

https://github.com/mrc-ide/COVID19_CFR_submission.

It consists of 436 patients from different countries (regions) which have records of three event time points: the time when symptoms appeared, the time when patient went to the hospital, and the time when the case was reported. The main survival time X^* is from symptom date to the date of reporting. This X^* is subject to left truncation by Y^*, the time from symptom date to the date of hospital admission (or testing date). We can only observe the pair (X^*, Y^*) when $X^* > Y^*$. Here X^* may be censored by C^* which is the last follow-up date, for example from symptom date to recovery or from symptom date to other events which result reporting not happening yet. Our target is to study the truncation probabilities α_k under different groups, in particular, testing the hypotheses in (1). Note that $\alpha_k = \mathbf{P}(X^*_{\cdot,k} \geq Y^*_{\cdot,k})$, thus a higher $\hat{\alpha}_k$ represents a lower probability of under-reporting $(1 - \alpha_k)$.

4.1. Under-reporting probabilities under different age groups or different countries

Firstly, using Equation (6), the estimated truncation probabilities $\hat{\alpha}_k$ under different age groups are calculated and listed in Table 4. Totally 4 age groups are considered here, i.e. $K = 4$. We use the test statistic W provided by (7) to test the hypotheses in (1). The alternative hypothesis means that as age increases the truncation probabilities go down, i.e. the under-reporting probabilities go higher for older people. The p-value of this test is 0.1446, hence H_0 cannot be rejected under a 5% significance level, which means that by considering the full data set, there is no significant difference in the under-reporting probabilities for different age groups. However, this result is based on all data from different countries (regions), which does not take into account the heterogeneity of countries (regions).

The underlying *country* (*region*) factor may also play an important role in this analysis. It is because of the heterogeneity of under-reporting across different countries (regions),

Table 4. The estimated truncation probabilities of different age groups.

Age group	0–30	31–50	51–75	76–100
k	1	2	3	4
n_k	58	141	209	28
$\hat{\alpha}_k$	0.4437	0.5120	0.5246	0.3181

Table 5. The estimated truncation probabilities of different regions.

Region group	Singapore	Japan	Hong Kong	Taiwan	Mainland China
k	1	2	3	4	5
n_k	74	131	72	15	108
$\hat{\alpha}_k$	0.7506	0.6692	0.4056	0.3881	0.2557

Table 6. The estimated truncation probabilities of different age groups.

	Mainland China			Hong Kong			Japan			Singapore		
Age group	k	$\hat{\alpha}_k$	n_k	k	$\hat{\alpha}_k$	n_k	k	$\hat{\alpha}_k$	n_k	k	$\hat{\alpha}_k$	n_k
0–30	1	0.2607	19	1	0.8000	7	1	0.6429	15	1	0.7419	11
31–50	2	0.1784	39	2	0.6482	15	2	0.8279	30	2	0.8051	34
51–75	3	0.2710	37	3	0.5900	43	3	0.6689	79	3	0.8530	28
76–100	4	0.4687	13	4	0.2500	7	4	0.5391	7	4	–	1

which is likely due to the different track-and-trace policies in each country (region). Therefore, we further implement an analysis by partition the data into subsets according to country (region) (Mainland China, Hong Kong, Taiwan, Japan and Singapore). The truncation probability estimates for different regions (not considering age effects) are shown in Table 5. It can be seen that the truncation probabilities of different region groups are quite different. Similarly as before, we use the test statistic W provided by (7) to test the hypotheses in (1). The p-value is less than 0.001 which shows that there is a very strong evidence to reject the NULL hypothesis, even under a 0.001 significance level. This means that before February 2020 Singapore had the lowest under-reporting probability (only about 25%), then it follows by Japan, Hong Kong, Taiwan. Mainland China had the highest under-reporting probability (nearly 75%).

4.2. Under-reporting probabilities under different age groups and regions

In this subsection, we will study the under-reporting probabilities under different age groups within each region. We will focus on Mainland China, Hong Kong, Japan, and Singapore. The data from Taiwan were dropped off because of not enough observations available. We display all the estimated truncation probabilities in Figure 1. It can be easily seen that in Hong Kong and Japan, the truncation probabilities tend to decrease with age increasing. However, in Mainland China and Singapore, the pattern is not clear or even seems to be opposite.

We perform the same hypothesis test for different age groups within each region, where the label information and partition details of each age groups are presented in Table 6. The p-values for Hong Kong and Japan are $9.0596 * 10^{-10}$ and 0.0287, respectively, which means the proportion of under-reporting for elder people is much larger than young people in these two areas. Meanwhile, the p-values for Mainland China and Singapore are 0.8631 and 0.7671, respectively. This means that in Mainland China and Singapore age is not a significant factor for under-reporting rates at a 5% significance level, although overall before February 2020 Mainland China had a very high under-reporting probability and Singapore had a very low under-reporting probability.

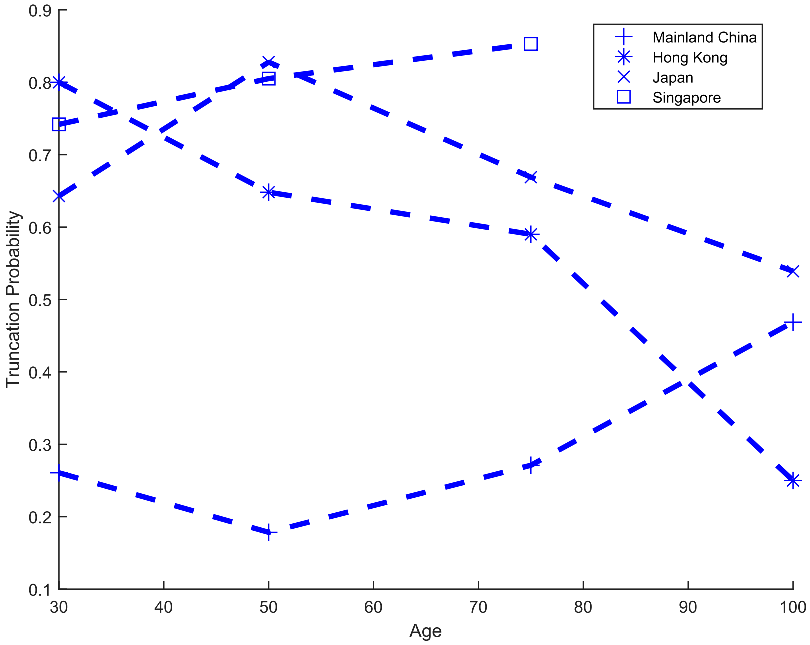

Figure 1. The estimated truncation probabilities of Mainland China, Hong Kong, Japan, and Singapore.

5. Conclusion

Under-reporting for COVID-19 infectious cases has become a major concern of World Health Organisation (WHO) since in order to better understand and control the pandemic, under-reporting probability should be lowered at a minimum possible level. In this study, we propose a truncation model to estimate the under-reporting probabilities for infected cases of COVID-19 in different countries and for different age groups. Our results largely agree with existing research. For example, the reports from Imperial College London, (Imai, Dorigatti, Cori, Riley, and Ferguson 2020b; Imai et al. 2020a), have pointed out the there could be massive under-reporting cases in China before February 2020. However, these existing studies use some very naive approaches or simulation studies to estimate the under-reporting probability and thus provide less reliable estimation. Existing methods may improve their estimates on under-reporting probability by using more extra information available only at very late stage of the pandemic. Therefore, for such fast spreading infectious disease, to better understand and control it, our proposed methodology gives an alternative sophisticated solution and can be used at very early stage of the pandemic (we only used some publicly available data from January to February 2020 but the method can still provide consistent estimate).

Our proposed methodology can provide consistent estimate for the truncation probability (thus for under-reporting probability) and can implement the hypothesis testing to find out which factors are related to under-reporting. First, the country (region) factor played an important role in the analysis, because of the heterogeneity of under-reporting across different countries. Second, we have found that in some countries, elder people had a significant higher under-reporting probability than younger people. One reason may be because many old people who live lonely or live in care may have died from the disease without even diagnosed. This was actually what happened in the UK and China, where both countries have once adjusted their total number of figures by adding the under-reporting figures in the later stage of the pandemic. This was evidenced by a single day spike of infected cases in both countries.

The disease has spread to almost every country in the world and the pandemic is still in the early stage in many countries. Therefore, our methods may be valuable for them to understand the epidemic, to bring about relevant actions to tackle it, to provide guidance on their decision makings and to control the epidemic and minimise its impacts on their economies and societies.

This research focussed on estimation and hypothesis testing for under-reporting probabilities based on univariate truncation models, where only hospital admission time and reporting time were used. Apart from these two events, the infected cases will have two outcomes in the end, cure or death. Therefore, a competing risk model can be used. The cure and death event will be subject to right censoring. Therefore, a future research work will be to study the under-reporting probabilities using bivariate truncation and censoring methodologies (Dai and Fu 2012; Dai, Restaino, and Wang 2016; Dai, Wang, Restaino, and Bao 2018; Wang, Dai, and Fu 2013), under the competing risk model framework.

Note

1. Note that the notations X and Y stand for the observable data, which must satisfy the condition $X \geq Y$, hence they are not independent. The independence assumption in Section 2.1 is for X^* and Y^* which can not be observed when $X^* < Y^*$.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Natural Science Foundation of China, Grant Number: 11701484; the Fundamental Research Funds for the Central Universities, Grant Number: 20720190067.

ORCID

Marialuisa Restaino http://orcid.org/0000-0002-1150-8278

References

Bilker, W.B., and Wang, M.C. (1996), ‘A Semiparametric Extension of the Mann-Whitney Test for Randomly Truncated Data’, *Biometrika*, 52, 10–20.

Dai, H., and Fu, B. (2012), ‘A Polar Coordinate Transformation for Estimating Bivariate Survival Functions with Randomly Censored and Truncated Data’, *Journal of Statistical Planning and Inference*, 142, 248–262.

Dai, H., Restaino, M., and Wang, H. (2016), ‘A Class of Nonparametric Bivariate Survival Function Estimators for Randomly Censored and Truncated Data’, *Journal of Nonparametric Statistics*, 28, 736–751.

Dai, H., Wang, H., Restaino, M., and Bao, Y. (2018), ‘Linear Transformation Models for Censored Data Under Truncation’, *Journal of Statistical Planning and Inference*, 193, 42–54.

Fleming, T.R., and Harrington, D.P. (1991), *Counting Processes and Survival Analysis*, Hoboken, NJ: John Wiley & Sons.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J., and Cao, B. (2020), 'Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China', *Lancet*, 395, 497–506. doi:10.1016/S0140-6736(20)30183-5.

Imai, N., Dorigatti, I., Cori, A., Riley, S., Donnelly, C., and Ferguson, N. (2020a), 'Report 2: Estimating the Potential Total Number of Novel Coronavirus (2019-nCoV) Cases in Wuhan City, China'.

Imai, N., Dorigatti, I., Cori, A., Riley, S., and Ferguson, N. (2020b), 'Report 1: Estimating the Potential Total Number of Novel Coronavirus (2019-nCoV) Cases in Wuhan City, China'.

Kalbfleisch, J.D., and Prentice, R.L. (2011), *The Statistical Analysis of Failure Time Data*, Hoboken, NJ: John Wiley & Sons.

Lagakos, S.W., Barraj, L.M., and De Gruttola, V. (1988), 'Nonparametric Analysis of Truncated Survival Data with Application to AIDS', *Biometrika*, 75, 515–524.

Lawless, J.F. (2003), *Statistical Models and Methods for Lifetime Data*, Hoboken, NJ: John Wiley & Sons.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S.M., Lau, E.H.Y., Wong, J.Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q., Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T.T.Y., Wu, J.T., Gao, G.F., Cowling, B.J., Yang, B., Leung, G.M., and Feng, Z. (2020), 'Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia', *The New England Journal of Medicine*, 382, 1199–1207. doi:10.1056/NEJMoa2001316.

Lynden-Bell, D. (1971), 'A Method for Allowing for Known Observational Selection in Small Samples Applied to 3CR Quasars', *Monthly Notices of the Royal Astronomical Society*, 155, 95–118.

Maugeri, A., Barchitta, M., Battiato, S., and Agodi, A. (2020), 'Estimation of Unreported Novel Coronavirus (SARS-CoV-2) Infections From Reported Deaths: A Susceptible-Exposed-Infectious-Recovered-Dead Model', *Journal of Clinical Medicine*, 9(5), 1350. doi:10.3390/jcm9051350.

Nishiura, H., Kinoshita, R., Yang, Y., Hayashi, K., Kobayashi, T., Yuan, B., and Akhmetzhanov, A.R. (2020), 'The Extent of Transmission of Novel Coronavirus in Wuhan, China', *Journal of Clinical Medicine*, 9, 330. doi:10.3390/jcm9020330.

Shen, P.S. (2007), 'A General Class of Test Procedures for Left-truncated and Right-censored Data', *Communications in Statistics: Theory Methods*, 36, 2913–2925.

Shen, P.S. (2009), 'A Class of Rank-based Test for Left-truncated and Right-censored Data', *Annals of the Institute of Statistical Mathematics*, 61, 461–476.

Shen, P.S. (2015), 'Nonparametric Tests for Left-truncated and Interval-censored Data', *Journal of Statistical Computation and Simulation*, 85, 1544–1553.

Tsai, W.Y. (1988), 'Estimation of the Survival Function with Increasing Failure Rate Based on Left Truncated and Right Censored Data', *Biometrika*, 75, 319–324.

Wang, H., Dai, H., and Fu, B. (2013), 'Accelerated Failure Time Models for Censored Survival Data Under Referral Bias', *Biostatistics*, 14, 313–326.

Woodroofe, M. (1985), 'Estimating a Distribution Function with Truncated Data', *Annals of Statistics*, 13, 163–177.

Zhang, Y. (2020), 'Analysis of Epidemiological Characteristics of New Coronavirus Pneumonia', *Chinese Journal of Epidemiology*, 41, 1–7.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., Chen, H.D., Chen, J., Luo, Y., Guo, H., Jiang, R.D., Liu, M.Q., Chen, Y., Shen, X.R., Wang, X., Zheng, X.S., Zhao, K., Chen, Q.J., Deng, F., Liu, L.L., Yan, B., Zhan, F.X., Wang, Y.Y., Xiao, G.F., and Shi, Z.L. (2020), 'A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin', *Nature*, 579, 270–273.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., Tan, W., and China Novel Coronavirus Investigating and Research Team (2020), 'A Novel Coronavirus From Patients with Pneumonia in China', *The New England Journal of Medicine*, 382, 727–733.

Appendices

Appendix 1. Large sample properties for $\hat{\Lambda}_k(t)$ and $\hat{S}_k(t)$

Define the filtration $\mathcal{F}_{i,k,t} = \sigma\{N_{i,k}(u), I[\tilde{X}_{i,k} \le t, \delta_{i,k} = 0],\ I[Y_{i,k} \le u] : 0 \le u \le t\},\ t \in [0, \tau]$, and denote its left limit as $\mathcal{F}_{i,k,t-}$. Then we have the following lemma.

Lemma A.1: *With the notations in Section 2, we have*

$$\mathbb{E}\{\mathrm{d}N_{i,k}(t) \mid \mathcal{F}_{i,k,t-}\} = I[\tilde{X}_{i,k} \ge t > Y_{i,k}]\, \mathrm{d}\Lambda_k(t) := \mathrm{d}A_{i,k}(t).$$

Proof: We need to show that for any set $B \in \mathcal{F}_{i,k,t-}$,

$$\mathbf{P}\left(\{\tilde{X}_{i,k} \ge t > Y_{i,k}\} \cap B\right) \cdot \mathrm{d}\Lambda_k(t) = \mathbf{P}\Big(\{\mathrm{d}N_{i,k}(t) = 1\} \cap B\Big). \qquad \text{(A1)}$$

For $B \in \mathcal{F}_{i,k,t-}$, there are the following three scenarios.

(a) A set $B \in \mathcal{F}_{i,k,t-}$, in the form of $\{\tilde{X}_{i,k} \ge s_3, Y_{i,k} \in [s_1, s_2)\}$ for any $0 \le s_1 < s_2 < t, s_3 \le t$, guarantees (A1) since it becomes

$$\mathbf{P}\left(\tilde{X}_{i,k} \ge t, Y_{i,k} \in [s_1, s_2)\right) \mathrm{d}\Lambda_k(t) = \mathbf{P}\left(\tilde{X}_{i,k} \in [t, t + \mathrm{d}t), \delta_{i,k} = 1, Y_{i,k} \in [s_1, s_2)\right),$$

or equivalently

$$\begin{aligned}&\mathbf{P}\left(\tilde{X}^*_{i,k} \ge t, Y^*_{i,k} \in [s_1, s_2) \mid X^*_{i,k} > Y^*_{i,k}\right) \mathrm{d}\Lambda_k(t)\\ &\quad = \mathbf{P}\left(\tilde{X}^*_{i,k} \in [t, t + \mathrm{d}t), \delta^*_{i,k} = 1, Y^*_{i,k} \in [s_1, s_2) \mid X^*_{i,k} > Y^*_{i,k}\right),\end{aligned}$$

which is obvious true.

(b) A set $B \in \mathcal{F}_{i,k,t-}$, in the form of $\{\tilde{X}_{i,k} \in [s_3, s_4), \delta_{i,k} = 1, Y_{i,k} \in [s_1, s_2)\}$ (or $\{\tilde{X}_{i,k} \in [s_3, s_4), \delta_{i,k} = 0, Y_{i,k} \in [s_1, s_2)\}$) for any $0 \le s_3 < s_4 \le t$ and any s_1, s_2 such that $B \in \mathcal{F}_{i,k,t-}$, also guarantees (A1) since both sides of the equation become 0.

(c) A set $B \in \mathcal{F}_{i,k,t-}$, in the form of $\{\tilde{X}_{i,k} \ge s_3, Y_{i,k} \ge s_1\}$ for any $0 \le s_1 \le t, 0 \le s_3 \le t$, guarantees (A1) since it becomes

$$\mathbf{P}\left(\tilde{X}_{i,k} \ge t > Y_{i,k} \ge s_1\right) \cdot \mathrm{d}\Lambda_k(t) = \mathbf{P}\left(\tilde{X}_{i,k} \in [t, t + \mathrm{d}t), \delta_{i,k} = 1, Y_{i,k} \ge s_1\right),$$

and further

$$\frac{\mathbf{P}\left(\tilde{X}^*_{i,k} \ge t, t > Y^*_{i,k} \ge s_1\right)}{\mathbf{P}\left(X^*_{i,k} > Y^*_{i,k}\right)} \cdot \mathrm{d}\Lambda_k(t) = \frac{\mathbf{P}\left(\tilde{X}^*_{i,k} \in [t, t + \mathrm{d}t), \delta^*_{i,k} = 1, t > Y^*_{i,k} \ge s_1\right)}{\mathbf{P}\left(X^*_{i,k} > Y^*_{i,k}\right)},$$

which is also true.

For any other sets, which are union, intersection, complementary of the sets of form in the above item (a), (b) and (c), Equation (A1) also holds. ■

Let $\mathrm{d}A_{i,k}(t) = I[\tilde{X}_{i,k} \ge t > Y_{i,k}]\, \mathrm{d}\Lambda_k(t)$ and $M_{i,k}(t) = N_{i,k}(t) - A_{i,k}(t)$. From Lemma A.1, we have $\mathbb{E}\{\mathrm{d}M_{i,k}(t) \mid \mathcal{F}_{i,k,t-}\} = 0$, i.e. $M_{i,k}(t)$ is a martingale with respect to $\mathcal{F}_{i,k,t}$. Similar arguments hold for $M_k(t) = \sum_{i=1}^{n_k} M_{i,k}(t)$ with respect to filtration $\mathcal{F}_{k,t} = \bigvee_{i=1}^{n_k} \mathcal{F}_{i,k,t}$.

From the definition of $\hat{\Lambda}_k$, we have

$$\hat{\Lambda}_k(t) - \Lambda_k(t) = \int_{[0,t]} \frac{I[\bar{H}_k(s) > 0]}{\bar{H}_k(s)}\, \mathrm{d}M_k(s) - \int_{[0,t]} I[\bar{H}_k(s) = 0]\, \mathrm{d}\Lambda_k(s),$$

we know that $\hat{\Lambda}_k(t)$ is a biased estimate to $\Lambda_k(t)$, but the bias $\int_{[0,t]} I[\bar{H}_k(s) = 0]\, \mathrm{d}\Lambda_k(s)$ (even with a multiple factor n_k) is negligible, according to Lemma A.3. Therefore, the variance of $\hat{\Lambda}_k(t)$ is such that

$$n_k \mathrm{Var}(\hat{\Lambda}_k(t)) = n_k \int_{[0,t]} \mathbb{E}\left\{\frac{I[\bar{H}_k(s) > 0]}{\bar{H}_k(s)}\right\} \mathrm{d}\Lambda_k(s) + o(1). \qquad \text{(A2)}$$

We also have the following corollary.

Corollary A.1: *For every $k = 1, 2, \dots, K$, under Conditions* 2.1, 2.2 *and Assumption* 2.1, *for $t \in [0, \tau_k]$, we have the following statements:*

$$\sqrt{n_k}\left(\hat{\Lambda}_k(t) - \Lambda_k(t)\right) \Rightarrow Z_{\Lambda_k}(t),$$

where $Z_{\Lambda_k}(t)$ is a zero-mean Gaussian process with independent increments and variance function

$$\sigma^2_{\Lambda_k}(t) = \alpha_k \int_0^t \frac{\mathrm{d}\Lambda_k(s)}{G_k(s)(1 - Q_k(s))S_k(s)}. \tag{A3}$$

Based on the following Lemma, which can be proved similarly as that in Fleming and Harrington (1991), we can have the large sample properties for $\hat{S}_k(t)$.

Lemma A.2: *If $S_k(t) > 0$,*

$$\frac{\hat{S}_k(t)}{S_k(t)} = 1 - \int_0^t \frac{\hat{S}_k(s-)}{S_k(s)} \left\{\frac{\mathrm{d}N_k(s)}{\bar{H}_k(s)} - \mathrm{d}\Lambda_k(s)\right\}.$$

The bias of $\hat{S}_k(t)$ is then given by

$$\hat{S}_k(t) - S_k(t) = -S_k(t) \int_0^t \frac{\hat{S}_k(s-)}{S_k(s)} \frac{I[\bar{H}_k(s) > 0]}{\bar{H}_k(s)}\,\mathrm{d}M_k(s) + B_k(t), \tag{A4}$$

where $B_k(t) = S_k(t) \int_0^t \frac{\hat{S}_k(s-)}{S_k(s)} I[\bar{H}_k(s) = 0]\,\mathrm{d}\Lambda_k(s)$. From Lemma A.3, $B_k(t)$ is negligible even with a multiplicative factor n_k. Hence, the variance of $\hat{S}_k(t)$ is such that

$$n_k \mathrm{Var}(\hat{S}_k(t)) = n_k S_k^2(t) \int_0^t \mathbb{E}\left\{\frac{\hat{S}_k^2(s-)}{S_k^2(s)} \frac{I[\bar{H}_k(s) > 0]}{\bar{H}_k(s)}\right\} \mathrm{d}\Lambda_k(s) + o(1). \tag{A5}$$

Corollary A.2: *Under Conditions* 2.1, 2.2 *and Assumption* 2.1, *for $t \in [0, \tau_k]$, we have the following statements:*

$$\sqrt{n_k}(\hat{S}_k(t) - S_k(t)) \Rightarrow Z_{S_k}(t),$$

where $Z_{S_k}(t)$ is a zero-mean Gaussian process with independent increments and variance function

$$\sigma^2_{S_k}(t) = \alpha_k\, S_k^2(t) \int_0^t \frac{\mathrm{d}\Lambda_k(s)}{G_k(s)(1 - Q_k(s))S_k(s)}. \tag{A6}$$

Appendix 2. Proof of Theorem 2.1

A.1. Necessary lemmas for proving Theorem 2.1

Lemma A.3: *Under Condition* 2.1 *and Assumption* 2.1, *we have the following statements: for any $\kappa \in (0, 1)$, $k = 1, 2, \dots, K$, we have*

$$n_k^{1-\kappa} \int_{[0,\tau_k]} I[\bar{H}_k(s) = 0]\,\mathrm{d}\Lambda_k(s) \xrightarrow{p} 0; \tag{A7}$$

Proof: According to Condition 2.1 we know that $\mathbf{P}(C_{i,k} \geq s > Y_{i,k}) \cdot S_k(s-) > 0$ with $s \in (0, \tau_k)$. Since $\mathbb{E}\, I[\bar{H}_k(s) = 0] = (1 - \mathbf{P}(C_{i,k} \geq s > Y_{i,k})S_k(s-))^{n_k}$, Assumption 2.1 implies that $\Lambda_k(\tau) < \infty$,

Λ_k is continuous at 0 and τ and

$$\int_{[0,\tau_k]} \Big(\mathbf{P}(C_{i,k} \geq s > Y_{i,k}) \cdot S_k(s-)\Big)^{-1} \,\mathrm{d}\Lambda_k(s) < \infty.$$

Therefore, we can find sequences $\epsilon_n \leq \epsilon$ and $\tau_n \geq \tau_0$, such that $\epsilon_n \downarrow 0$ and $\tau_n \uparrow \tau_k$ as $n \to \infty$ and $\Lambda_k = 1/({n_k}^{1-\kappa} \log n_k)$, $\Lambda_k(\tau_k) - \Lambda_k(\tau_n) = 1/(n_k^{1-\kappa} \log n_k)$. Because

$$\int_{[0,\tau_k]} \Big(\mathbf{P}(C_{i,k} \geq s > Y_{i,k}) \cdot S_k(s-)\Big)^{-1} \,\mathrm{d}\Lambda_k(s) < \infty,$$

we know that

$$\Big(\mathbf{P}(C_{i,k} \geq \epsilon_n > Y_{i,k}) \cdot S_k(\epsilon_n-)\Big)^{-1} \cdot \Lambda_k(\epsilon_n) \leq \int_{[0,\epsilon_n]} \Big(\mathbf{P}(C_{i,k} \geq s > Y_{i,k}) \cdot S_k(s-)\Big)^{-1} \,\mathrm{d}\Lambda_k(s) < \infty$$

and further

$$\mathbf{P}(C_{i,k} \geq \epsilon_n > Y_{i,k}) \cdot S_k(\epsilon_n-) \geq \Lambda_k(\epsilon_n) = \frac{1}{n_k^{1-\kappa} \log n_k}.$$

Similarly we have $\mathbf{P}(C_{i,k} \geq \tau_n > Y_{i,k}) \cdot S_k(\tau_k-) \geq (n_k^{1-\kappa} \log n_k)^{-1}$.

On the other hand, $\mathbf{P}(C_{i,k} \geq s > Y_{i,k}) \cdot S_k(s-) = \mathbb{E}(n_k^{-1}\bar{H}_k(s))$, and

$$\begin{aligned}
&\mathbb{E}\left| n_k^{1-\kappa} \int_{[0,\tau_k]} I[\bar{H}_k(s) = 0] \,\mathrm{d}\Lambda_k(s) \right| = n_k^{1-\kappa} \int_{[0,\tau_k]} [1 - \mathbb{E}(n_k^{-1}\bar{H}_k(s))]^{n_k} \,\mathrm{d}\Lambda_k(s) \\
&= n_k^{1-\kappa} \int_{[0,\epsilon_n]} \big[1 - \mathbb{E}(n_k^{-1}\bar{H}_k(s))\big]^{n_k} \,\mathrm{d}\Lambda_k(s) + n_k^{1-\kappa} \int_{[\tau_n,\tau_k]} [1 - \mathbb{E}(n_k^{-1}\bar{H}_k(s))]^{n_k} \,\mathrm{d}\Lambda_k(s) \\
&\quad + n_k^{1-\kappa} \int_{[\epsilon_n,\tau_n]} [1 - \mathbb{E}(n_k^{-1}\bar{H}_k(s))]^{n_k} \,\mathrm{d}\Lambda_k(s) \\
&\leq n_k^{1-\kappa}\, \Lambda_k(\epsilon_n) + n_k^{1-\kappa}\, (\Lambda_k(\tau_k) - \Lambda_k(\tau_n)) + n_k^{1-\kappa} [1 - 1/(n_k^{1-\kappa} \log n_k)]^{n_k}\, \Lambda_k(\tau_k) \\
&= \frac{2}{\log n_k} + O\big(n_k^{1-\kappa} \exp(-n_k^{\kappa}/\log n_k)\big) \to 0
\end{aligned}$$

which implies (A7). The lemma is proved. ■

Let $\mathrm{d}N_{k,G}(t) = \sum_{i=1}^{n_k} I[t \leq Y_{i,k} < t + \mathrm{d}t]$, $\mathrm{d}A_{k,G}(t) = \sum_{i=1}^{n_k} I[Y_{i,k} \geq t]\mathrm{d}\Lambda_{k,G}(t)$. Similarly as Lemma A.1, we have that $M_{k,G}(t) = N_{k,G}(t) - A_{k,G}(t)$ is a martingale with respect to $\mathcal{F}_{k,t}$ and we also have the following lemmas.

Lemma A.4: *If* $G_k(t) > 0$,

$$\frac{\hat{G}_k(t)}{G_k(t)} = 1 - \int_t^{\tau_k} \frac{\hat{G}_k(s-)I[\bar{H}_k(s) > 0]}{G_k(s)\bar{H}_k(s)} \left(\mathrm{d}M_{k,G}(s) + \frac{J_k(s)}{G_k(s)}\mathrm{d}\Lambda_{k,G}(s)\right) + \Delta_{G_k}, \qquad \text{(A8)}$$

where $J_k(s) = \bar{H}_k(s) - G_k(s) \sum_{i=1}^{n_k} I[\tilde{X}_{i,k} \geq s] = O_p(n_k^{-1/2})$, *and*

$$\Delta_{G_k} = -\int_t^{\tau_k} \frac{\hat{G}_k(s-)I[\bar{H}_k(s) = 0]}{G_k(s)} \,\mathrm{d}\Lambda_{k,G}(s).$$

Proof: Suppose U, V and W are right-continuous functions of locally bounded variation on any finite interval $[t, \tau_k]$, then $U(\tau_k)V(\tau_k) - U(t)V(t) = \int_t^{\tau_k} U(s-)\,\mathrm{d}V(s) + \int_t^{\tau_k} V(s)\,\mathrm{d}U(s)$. and

$\mathrm{d}W^{-1}(s) = -(W(s)W(s-))^{-1}\mathrm{d}W(s)$. Let $U(s) = \hat{G}_k(s)$, $W(s) = G_k(s)$, $V(s) = G_k^{-1}(s)$, so

$$\frac{\hat{G}_k(t)}{G_k(t)} = 1 + \int_t^{\tau_k} \frac{\hat{G}_k(s-)}{G_k(s)} \frac{\mathrm{d}G_k(s)}{G_k(s)} - \int_t^{\tau_k} \frac{\hat{G}_k(s-)}{G_k(s)} \frac{\mathrm{d}\hat{G}_k(s)}{\hat{G}_k(s-)}.$$

From Theorem 1 in Woodroofe (1985), $\hat{G}_k^{-1}(s-)\,\mathrm{d}\hat{G}_k(s) = \bar{H}_k^{-1}(s)\,\mathrm{d}N_{k,G}(s)$, together with $\mathrm{d}G_k(s) = (1 - G_k(s))\,\mathrm{d}\Lambda_{k,G}(s)$, we get

$$\begin{aligned}
\frac{\hat{G}_k(t)}{G_k(t)} &= 1 - \int_t^{\tau_k} \frac{\hat{G}_k(s-)I[\bar{H}_k(s) > 0]}{G_k(s)\bar{H}_k(s)} \mathrm{d}M_{k,G}(s) \\
&\quad + \int_t^{\tau_k} \frac{\hat{G}_k(s-)I[\bar{H}_k(s) > 0]}{G_k^2(s)\bar{H}_k(s)} \left(\bar{H}_k(s) - G_k(s) \sum_{i=1}^{n_k} I[\tilde{X}_{i,k} \geq s] \right) \mathrm{d}\Lambda_{k,G}(s) \\
&\quad - \int_t^{\tau_k} \frac{\hat{G}_k(s-)I[\bar{H}_k(s) = 0]}{G_k(s)} \mathrm{d}\Lambda_{k,G}(s) \\
&= 1 - \int_t^{\tau_k} \frac{\hat{G}_k(s-)I[\bar{H}_k(s) > 0]}{G_k(s)\bar{H}_k(s)} \left(\mathrm{d}M_{k,G}(s) + \frac{J_k(s)}{G_k(s)} \mathrm{d}\Lambda_{k,G}(s) \right) + \Delta_{G_k},
\end{aligned}$$

where $J_k(s) = \bar{H}_k(s) - G_k(s) \sum_{i=1}^{n_k} I[\tilde{X}_{i,k} \geq s]$, and

$$\Delta_{G_k} = - \int_t^{\tau_k} \frac{\hat{G}_k(s-)I[\bar{H}_k(s) = 0]}{G_k(s)} \mathrm{d}\Lambda_{k,G}(s).$$

Denote $L_k(s) = \mathbf{P}(\tilde{X}_{i,k} \geq s > Y_{i,k})$. Since

$$L_k(s) = \alpha_k^{-1} \mathbf{P}\left(X_{i,k}^* \geq s, C_{i,k}^* \geq s > Y_{i,k}^* \right) = \alpha_k^{-1} S_k(s)(1 - Q_k(s))G_k(s), \tag{A9}$$

and $G_k(s)\,\mathbf{P}(\tilde{X}_{i,k} \geq s) = G_k(s)\,\mathbf{P}(\min\{X_{i,k}, C_{i,k}\} \geq s) = \alpha_k^{-1} S_k(s)(1 - Q_k(s))G_k(s)$, we get $n_k^{-1} J_k(s) = n_k^{-1}\bar{H}_k(s) - G_k(s)(n_k^{-1} \sum_{i=1}^{n_k} I[\tilde{X}_{i,k} \geq s]) \to L_k(s) - G_k(s)\,\mathbf{P}(\tilde{X}_{i,k} \geq s) = 0$, and furthermore, from central limit theorem, we have $J_k(s) = O_p(n_k^{-1/2})$.

Since Δ_G is negligible (even with a multiple factor n_k) according to Lemma A.3, we can rewrite $\Delta_G = o_p(n_k^{-1/2})$, hence $\hat{G}_k$ is asymptotically unbiased, and

$$\hat{G}_k(t) - G_k(t) = -G_k(t) \int_t^{\tau_k} \frac{\hat{G}_k(s-)I[\bar{H}_k(s) > 0]}{G_k(s)\bar{H}_k(s)} \left(\mathrm{d}M_{k,G}(s) + J_k(s) \frac{\mathrm{d}\Lambda_{k,G}(s)}{G_k(s)} \right) + \bar{B}_{k,G}(t), \tag{A10}$$

where $\bar{B}_{k,G}(t) = G_k(t) \cdot \Delta_{G_k} = o_p(n_k^{-1/2})$. ■

Lemma A.5: *Under Conditions 2.1, 2.2 and Assumption 2.2, for $k = 1, 2, \ldots, K$, as $n \to \infty$,*

$$\sqrt{n} \int_0^{\tau_k} \left(\int_0^u G_k(s)\,\mathrm{d}S_k(s) \right) \frac{\hat{G}_k(u-)}{G_k^2(u)} \frac{I[\bar{H}_k(u) > 0]J_k(u)}{\bar{H}_k(u)} \mathrm{d}\Lambda_{k,G}(u) \to N\left(0, p_k^{-1}\sigma_{k3}^2\right),$$

where

$$\sigma_{k3}^2 = 2 \int_0^{\tau_k} \left(\int_0^v \frac{\int_0^u G_k(s)\,\mathrm{d}S_k(s)}{L_k(u)} \mathrm{d}\Lambda_{k,G}(u) \right) \frac{\int_0^v G_k(s)\,\mathrm{d}S_k(s)}{G_k^2(v)} \mathrm{d}G_k(v).$$

Proof: Denote

$$W_{k,G}(u) = \left(\int_0^u G_k(s)\,\mathrm{d}S_k(s)\right)\left(\frac{\hat{G}_k(u-) - G_k(u)}{G_k(u)} + 1\right)\frac{I[\bar{H}_k(u) > 0]}{\bar{H}_k(u)}$$
$$= \left(\int_0^u G_k(s)\,\mathrm{d}S_k(s)\right)\frac{I[\bar{H}_k(u) > 0]}{\bar{H}_k(u)} + O_p(n_k^{-1/2}). \tag{A11}$$

Since $n_k^{-1}J_k(u) = O_p(n_k^{-1/2})$, and

$$\frac{1}{n_k^{-1}\bar{H}_k(u)} = \frac{1}{L_k(u)}\left(1 + \frac{L_k(u) - n_k^{-1}\bar{H}_k(u)}{n_k^{-1}\bar{H}_k(u)}\right),$$

where $L_k(u)$ is defined in (A9), therefore

$$\int_0^{\tau_k}\frac{W_{k,G}(u)\,J_k(u)}{G_k(u)}\mathrm{d}\Lambda_{k,G}(u) \tag{A12}$$
$$= \int_0^{\tau_k}\left(\int_0^u G_k(s)\,\mathrm{d}S_k(s)\right)\frac{I[\bar{H}_k(u) > 0]n_k^{-1}J_k(u)}{G_k(u)}\frac{1}{n_k^{-1}\bar{H}_k(u)}\mathrm{d}\Lambda_{k,G}(u) + O_p(n_k^{-1})$$
$$= \int_0^{\tau_k}\left(\int_0^u G_k(s)\,\mathrm{d}S_k(s)\right)\frac{I[\bar{H}_k(u) > 0]n_k^{-1}J_k(u)}{G_k(u)L_k(u)}\left(1 + \frac{L_k(u) - n_k^{-1}\bar{H}_k(u)}{n_k^{-1}\bar{H}_k(u)}\right)$$
$$\times\,\mathrm{d}\Lambda_{k,G}(u) + O_p(n_k^{-1})$$
$$= \int_0^{\tau_k}\left(\int_0^u G_k(s)\,\mathrm{d}S_k(s)\right)\frac{n_k^{-1}J_k(u)}{G_k(u)L_k(u)}\mathrm{d}\Lambda_{k,G}(u) + O_p(n_k^{-1})$$
$$= \frac{1}{n_k}\sum_{i=1}^{n_k}\int_0^{\tau_k}\frac{\int_0^u G_k(s)\,\mathrm{d}S_k(s)}{G_k(u)L_k(u)}\left(I[\tilde{X}_{i,k} \ge u > Y_{i,k}] - G_k(u)I[\tilde{X}_{i,k} \ge u]\right)\mathrm{d}\Lambda_{k,G}(u) + O_p(n_k^{-1}). \tag{A13}$$

Let $J_{k,i}(u) = I[\tilde{X}_{i,k} \ge u > Y_{i,k}] - G_k(u)I[\tilde{X}_{i,k} \ge u]$, and

$$\xi_{i,k} = \int_0^{\tau_k}\frac{\int_0^u G_k(s)\,\mathrm{d}S_k(s)}{G_k(u)L_k(u)}J_{k,i}(u)\,\mathrm{d}\Lambda_{k,G}(u),$$

then $\mathbb{E}\,\xi_{i,k} = 0$. Denote

$$\sigma_{k3}^2 = \mathbb{E}\,\xi_{i,k}^2 = \mathbb{E}\left(\int_0^{\tau_k}\int_0^{\tau_k}\frac{\int_0^u G_k(s)\,\mathrm{d}S_k(s)}{G_k(u)\,L_k(u)}\frac{\int_0^v G_k(s)\,\mathrm{d}S_k(s)}{G_k(v)\,L_k(v)}J_{k,i}(u)\,J_{k,i}(v)\,\mathrm{d}\Lambda_{k,G}(u)\,\mathrm{d}\Lambda_{k,G}(v)\right).$$

Since

$$\mathbb{E}(J_{k,i}(u)J_{k,i}(v)) = \mathbb{E}\Big(I[\tilde{X}_{i,k} \ge u]\,I[\tilde{X}_{i,k} \ge v]\left(I[Y_{i,k} < u] - G_k(u)\right)\left(I[Y_{i,k} < v] - G_k(v)\right)\Big)$$
$$= \alpha_k^{-1}\mathbb{E}\left(I[\tilde{X}_{i,k}^* \ge \max(u,v)]\left(I[Y_{i,k}^* < u] - G_k(u)\right)\left(I[Y_{i,k}^* < v] - G_k(v)\right)\right)$$
$$= \alpha_k^{-1}S_k(\max(u,v))\Big(1 - Q_k(\max(u,v))\Big)\Big(G_k(\min(u,v)) - G_k(u)G_k(v)\Big),$$

together with the definition of $L_k(\cdot)$ in (A9), we have

$$\sigma_{k3}^2 = \iint_{0\le u\le v\le\tau_k}\frac{\int_0^u G_k(s)\,\mathrm{d}S_k(s)}{L_k(u)}\frac{\int_0^v G_k(s)\,\mathrm{d}S_k(s)}{G_k^2(v)}(1 - G_k(v))\,\mathrm{d}\Lambda_{k,G}(u)\,\mathrm{d}\Lambda_{k,G}(v)$$
$$+ \iint_{\tau_k\ge u>v\ge 0}\frac{\int_0^u G_k(s)\,\mathrm{d}S_k(s)}{G_k^2(u)}\frac{\int_0^v G_k(s)\,\mathrm{d}S_k(s)}{L_k(v)}(1 - G_k(u))\,\mathrm{d}\Lambda_{k,G}(u)\,\mathrm{d}\Lambda_{k,G}(v)$$

$$= 2 \iint_{0 \le u \le v \le \tau_k} \frac{\int_0^u G_k(s)\, \mathrm{d}S_k(s)}{L_k(u)} \frac{\int_0^v G_k(s)\, \mathrm{d}S_k(s)}{G_k^2(v)} \, \mathrm{d}\Lambda_{k,G}(u)\, \mathrm{d}G_k(v).$$

Notice that, under Assumption 2.2,

$$\sigma_{k3}^2 < 2 \iint_{0 \le u \le v \le \tau_k} \frac{\mathrm{d}\Lambda_{k,G}(u)\, \mathrm{d}G_k(v)}{L_k(u)\, G_k^2(v)} < 2 \int_0^{\tau_k} \left(\frac{1}{G_k(u)} - 1 \right) \frac{\mathrm{d}\Lambda_{k,G}(u)}{L_k(u)} < \infty.$$

The lemma is proved. ■

A.2. Proof of Theorem 2.1

Proof: Under H_0, we have

$$W = \sum_{k=1}^{K} c_k \left\{ \int_0^{\tau_k} \hat{G}_k(s)\hat{S}_k(s) \left(\frac{\mathrm{d}N_k(s)}{\bar{H}_k(s)} - I[\bar{H}_k(s) > 0]\mathrm{d}\Lambda_k(s) - I[\bar{H}_k(s) = 0]\mathrm{d}\Lambda_k(s) \right) \right\}$$
$$+ \sum_{k=1}^{K} c_k \left\{ \int_0^{\tau_k} \left(\hat{G}_k(s)\hat{S}_k(s) - G_k(s)S_k(s) \right) \mathrm{d}\Lambda_k(s) \right\}$$
$$= \sum_{k=1}^{K} c_k \int_0^{\tau_k} \hat{G}_k(s)\hat{S}_k(s) \frac{I[\bar{H}_k(s) > 0]}{\bar{H}_k(s)} \mathrm{d}M_k(s) + \sum_{k=1}^{K} c_k \int_0^{\tau_k} \left(\hat{G}_k(s)\hat{S}_k(s) - G_k(s)S_k(s) \right) \mathrm{d}\Lambda_k(s)$$
$$- \sum_{k=1}^{K} c_k \int_0^{\tau_k} \hat{G}_k(s)\hat{S}_k(s) I[\bar{H}_k(s) = 0]\, \mathrm{d}\Lambda_k(s) = \sum_{k=1}^{K} c_k (U_{k1} + U_{k2} + U_{k3})$$

Since U_{k1} is a typical martingale and U_{k3} is negligible according to Lemma A.3, we only need to examine the properties of U_{k2}. Denote

$$U_{k2,S} = \int_0^{\tau_k} \left(\hat{S}_k(s) - S_k(s) \right) \hat{G}_k(s)\, \mathrm{d}\Lambda_k(s), \quad U_{k2,G} = \int_0^{\tau_k} \left(\hat{G}_k(s) - G_k(s) \right) S_k(s)\, \mathrm{d}\Lambda_k(s),$$

and then $U_{k2} = U_{k2,S} + U_{k2,S}$. Substituting equations (A4) and (A8) into $U_{k2,S}$ and $U_{k2,G}$, together with Lemma A.3, we get

$$U_{k2,S} = \int_0^{\tau_k} \left(-S_k(s) \int_0^s \frac{\hat{S}_k(u-)}{S_k(u)} \frac{I[\bar{H}_k(u) > 0]}{\bar{H}_k(u)} \mathrm{d}M_k(u) + B_k(s) \right) \hat{G}_k(s)\, \mathrm{d}\Lambda_k(s)$$
$$= \int_0^{\tau_k} \left(\int_u^{\tau_k} \hat{G}_k(s)\, \mathrm{d}S_k(s) \right) \frac{\hat{S}_k(u-)}{S_k(u)} \frac{I[\bar{H}_k(u) > 0]}{\bar{H}_k(u)} \mathrm{d}M_k(u) + o_p(n_k^{-1/2}),$$

and

$$U_{k2,G} = -\int_0^{\tau_k} G_k(s) \int_s^{\tau_k} \frac{\hat{G}_k(u-) I[\bar{H}_k(u) > 0]}{G_k(u)\bar{H}_k(u)}$$
$$\times \left(\mathrm{d}M_{k,G}(u) + J_k(u) \frac{\mathrm{d}\Lambda_{k,G}(u)}{G_k(u)} \right) S_k(s)\, \mathrm{d}\Lambda_k(s) + o_p(n_k^{-1/2})$$
$$= \int_0^{\tau_k} \left(\int_0^u G_k(s)\, \mathrm{d}S_k(s) \right) \frac{\hat{G}_k(u-)}{G_k(u)} \frac{I[\bar{H}_k(u) > 0]}{\bar{H}_k(u)}$$
$$\times \left(\mathrm{d}M_{k,G}(u) + J_k(u) \frac{\mathrm{d}\Lambda_{k,G}(u)}{G_k(u)} \right) + o_p(n_k^{-1/2}).$$

Hence $U_{k1}+U_{k2}=U_{k1}+U_{k2,S}+U_{k2,G}$ can be rewritten as $\mathrm{I}_k+\mathrm{II}_k+\mathrm{III}_k+o_p(n_k^{-1/2})$, where

$$
\begin{aligned}
\mathrm{I}_k &= \int_0^{\tau_k} W_k(u)\,\mathrm{d}M_k(u), \quad \mathrm{II}_k = \int_0^{\tau_k} W_{k,G}(u)\,\mathrm{d}M_{k,G}(u),\\
\mathrm{III}_k &= \int_0^{\tau_k} \frac{W_{k,G}(u)J_k(u)}{G_k(u)}\,\mathrm{d}\Lambda_{k,G}(u),\\
W_k(u) &= \left(\hat{G}_k(u)\hat{S}_k(u) + \left(\int_u^{\tau_k} \hat{G}_k(s)\,\mathrm{d}S_k(s)\right)\frac{\hat{S}_k(u-)}{S_k(u)}\right)\frac{I[\bar{H}_k(u)>0]}{\bar{H}_k(u)}\\
&= \left(\int_u^{\tau_k} S_k(s)\,\mathrm{d}\hat{G}_k(s)\right)\frac{\hat{S}_k(u-)}{S_k(u)}\frac{I[\bar{H}_k(u)>0]}{\bar{H}_k(u)},
\end{aligned}
$$

with III_k given in (A13) and $W_{k,G}(u)$ given in (A11).

Since both $M_k(t)$ and $M_{k,G}(t)$ are martingale, together with the definition of L_k in (A9), we have

$$
\begin{aligned}
\left\langle\sqrt{n_k}\,\mathrm{I}_k,\sqrt{n_k}\,\mathrm{I}_k\right\rangle &= n_k\int_0^{\tau_k} W_k^2(u)\,\mathrm{d}\left\langle M_k(u),M_k(u)\right\rangle = n_k\int_0^{\tau_k} W_k^2(u)\bar{H}_k(u)\,\mathrm{d}\Lambda_k(u)\\
&\to \int_0^{\tau_k}\left(\int_u^{\tau_k} S_k(s)\,\mathrm{d}G_k(s)\right)^2\frac{\mathrm{d}\Lambda_k(u)}{L_k(u)} := \sigma_{k1}^2,\\
\left\langle\sqrt{n_k}\,\mathrm{II}_k,\sqrt{n_k}\,\mathrm{II}_k\right\rangle &= n_k\int_0^{\tau_k} W_{k,G}^2(u)\,\mathrm{d}\left\langle M_{k,G}(u),M_{k,G}(u)\right\rangle\\
&= n_k\int_0^{\tau_k} W_{k,G}^2(u)\left(\sum_{i=1}^{n_k} I[\tilde{X}_{i,k}\ge Y_{i,k}\ge u]\right)\mathrm{d}\Lambda_{k,G}(u)\\
&\to \int_0^{\tau_k}\left(\int_0^u G_k(s)\,\mathrm{d}S_k(s)\right)^2\frac{\mathrm{d}G_k(u)}{\alpha_k L_k^2(u)} := \sigma_{k2}^2.
\end{aligned}
$$

Under Assumptions 2.1 and 2.2, both σ_{k1}^2 and σ_{k2}^2 exist.

From Theorem 2.6.1 in Fleming and Harrington (1991), we know

$$
\left\langle M_k(u),M_{k,G}(u)\right\rangle = -\int_0^u \Delta A_k(t)\,\mathrm{d}A_{k,G}(t) = -\sum_{i=1}^{n_k}\int_0^u \Delta A_k(t)\,I[\tilde{X}_{i,k}\ge Y_{i,k}\ge t]\,\mathrm{d}\Lambda_{k,G}(t).
$$

Further, since $\mathrm{d}A_k(t)=\sum_{i=1}^{n_k} I[\tilde{X}_{i,k}\ge t> Y_{i,k}]\,\mathrm{d}\Lambda_k(t)$, we get

$$
\left\langle\sqrt{n_k}\,\mathrm{I}_k,\sqrt{n_k}\,\mathrm{II}_k\right\rangle = n_k\int_0^{\tau_k} W_k(u)W_{k,G}(u)\,\mathrm{d}\left\langle M_k(u),M_{k,G}(u)\right\rangle = 0.
$$

According to the martingale theory in Fleming and Harrington (1991), under Assumptions *2.1* and *2.2*, when $n\to\infty$, $\sqrt{\frac{n}{n_k}}\sqrt{n_k}(\mathrm{I}_k+\mathrm{II}_k)\to N(0,p_k^{-1}(\sigma_{k1}^2+\sigma_{k2}^2))$. By Lemma A.5, we also have $\sqrt{\frac{n}{n_k}}\sqrt{n_k}\,\mathrm{III}_k\to N(0,p_k^{-1}\sigma_{k3}^2)$.

Therefore, using the result

$$
\begin{aligned}
&\mathrm{Cov}\Big((\mathrm{I}_k+\mathrm{II}_k),\mathrm{III}_k\Big) = \mathbb{E}\,(\mathrm{I}_k+\mathrm{II}_k)\mathrm{III}_k\\
&= \mathbb{E}\int_0^{\tau_k} W_k(u)\,\mathrm{d}M_k(u)\int_0^{\tau_k} W_{k,G}(v)J_k(v)\frac{\mathrm{d}\Lambda_{k,G}(v)}{G_k(v)}\\
&\quad+\mathbb{E}\int_0^{\tau_k} W_{k,G}(u)\,\mathrm{d}M_{k,G}(u)\int_0^{\tau_k} W_{k,G}(v)J_k(v)\frac{\mathrm{d}\Lambda_{k,G}(v)}{G_k(v)}
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left(\int_0^{\tau_k} W_k(u)\left(\int_0^{\tau_k} W_{k,G}(v)J_k(v)\frac{\mathrm{d}\Lambda_{k,G}(v)}{G_k(v)}\right)\mathrm{d}M_k(u)\right)\\
&\quad + \mathbb{E}\left(\int_0^{\tau_k} W_{k,G}(u)\left(\int_0^{\tau_k} W_{k,G}(v)J_k(v)\frac{\mathrm{d}\Lambda_{k,G}(v)}{G_k(v)}\right)\mathrm{d}M_{k,G}(u)\right) = 0,
\end{aligned}
$$

we get $\sqrt{\frac{n}{n_k}}\sqrt{n_k}(U_{k1}+U_{k2}+U_{k3}) \rightarrow N(0, p_k^{-1}\sigma_k^2)$, where $\sigma_k^2 = \sigma_{k1}^2 + \sigma_{k2}^2 + \sigma_{k3}^2$.

Since K groups are independent, then under H_0, we have

$$
\sqrt{n}\,W = \sum_{k=1}^{K} c_k\sqrt{n}(U_{k1}+U_{k2}+U_{k3}) \rightarrow N\left(0, \sigma_W^2\right),
$$

where $\sigma_W^2 = \sum_{k=1}^{K} p_k^{-1}c_k^2\sigma_k^2$. ■